frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 22 August 2025
DOI 10.3389/fcomp.2025.1629725

OPEN ACCESS

EDITED BY
Cenk Demiroglu,
Özyeğin University, Türkiye

REVIEWED BY
Bochao Zou,
University of Science and Technology Beijing, China
Biman Najika Liyanage,
Biman Liyanage, China

*CORRESPONDENCE
Yupei Li
✉ yl7622@ic.ac.uk

†These authors have contributed equally to this work

RECEIVED 16 May 2025
ACCEPTED 07 August 2025
PUBLISHED 22 August 2025

CITATION
Li Y, Shao S, Milling M and Schuller BW (2025) Large language models for depression recognition in spoken language integrating psychological knowledge.
Front. Comput. Sci. 7:1629725.
doi: 10.3389/fcomp.2025.1629725

Large language models for depression recognition in spoken language integrating psychological knowledge

Yupei Li[1*†], Shuaijie Shao[2†], Manuel Milling[3,4] and Björn W. Schuller[1,3,4,5]

[1]Group on Language, Audio, and Music, Imperial College London, London, United Kingdom, [2]University College London, London, United Kingdom, [3]Chair of Health Informatics, TUM University Hospital, Munich, Germany, [4]Munich Center for Machine Learning, Munich, Germany, [5]Munich Data Science Institute, Munich, Germany

Depression is a growing concern gaining attention in both public discourse and AI research. While deep neural networks (DNNs) have been used for its recognition, they still lack real-world effectiveness. Large language models (LLMs) show strong potential but require domain-specific fine-tuning and struggle with non-textual cues. Since depression is often expressed through vocal tone and behavior rather than explicit text, relying on language alone is insufficient. Diagnostic accuracy also suffers without incorporating psychological expertise. To address these limitations, we present, to the best of our knowledge, the first application of LLMs to multimodal depression detection using the DAIC-WOZ dataset. We extract the audio features using the pre-trained model Wav2Vec, and map them to text-based LLMs for further processing. We also propose a novel strategy for incorporating psychological knowledge into LLMs to enhance diagnostic performance, specifically using a question and answer set to grant authorized knowledge to LLMs. Our approach yields a notable improvement in both Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) compared to a base score proposed by the related original paper. The codes are available in Github.

KEYWORDS

large language models, depression recognition, psychological knowledge, spoken language, speech

1 Introduction

As mental health gains increasing attention from the public, the diagnosis of emotional disorders—particularly depression—has become an essential area in both AI for healthcare and medical research. In previous decades, diagnosis has solely depended on the expertise of professional psychologists and clinicians. These circumstances introduce potential variability, as different clinicians may be influenced by their own intuition when assessing subjective questions, evident by Gerber et al. (1989). To counteract this issue, diagnostic procedures are established with a maximum amount of standardization, which stays limited despite major efforts. One opportunity to increase consistency and reproducibility of diagnosis is by leveraging automated assessments as a supporting tool, which is enabled by large-scale data and AI techniques (Zafar et al., 2024). A combination of AI, vast experiential data, and the professional knowledge of clinicians may together contribute to more accurate depression recognition, enhancing clinicians' confidence and convenience,

as well as increasing patient satisfaction, which is shown in the work of Qassim et al. (2023). Although the use of AI alone to recognize depression has been proposed as a way to save time and reduce costs, patients have expressed skepticism and distrust, which Robertson et al. (2023) have revealed. Challenges include not only suboptimal detection rates but also the fact that current AI systems often lack integration with professional psychological knowledge, relying instead solely on data-driven experience.

Previous literature has made tremendous efforts to recognize depression within the domain of AI in healthcare. Two major research directions include the utilization and adaptation of various deep neural networks (DNNs), as well as the fusion of multiple data modalities. A wide range of techniques for depression detection has been surveyed by Squires et al. (2023). Notably, Valstar et al. (2013) introduced the depression detection challenge in 2013, laying the groundwork for subsequent research. Building on this, Ringeval et al. (2019) proposed the use of Long Short-Term Memory (LSTM) models for the task, establishing a baseline for future approaches. Since then, various deep neural network (DNN) models have been developed, achieving promising results. For instance, Niu et al. (2022) introduced Dual Attention and Element Recalibration Networks utilizing acoustic modalities for depression recognition. Despite their contributions, these models exhibit performance limitations largely due to the constrained learning capacity of their architectures. Recently, LLMs have demonstrated exceptional learning capabilities across many different fields (e. g., Amin et al., 2023); however, their potential has not yet been effectively explored in the context of depression detection. Additionally, unimodal approaches are often insufficient for accurately assessing depression, as the evaluation process is inherently complex. Even clinicians frequently rely on indirect questioning during interviews to avoid triggering sensitive responses. Some prior studies have explored modality-specific models, such as LLM-based models focused on text (Farruque et al., 2024), autoencoder-based models centered on audio features (Sardari et al., 2022), and models targeting image and video modalities (Ashraf et al., 2020). These approaches, however, still lack definitive and comprehensive diagnostic information due to their unimodal nature. For instance, while audio may reveal weak pitch variations, the transcribed textual content might appear entirely normal, obscuring underlying emotional cues. Consequently, multimodal fusion is essential for a more holistic understanding of depressive symptoms. However, integrating multiple modalities into LLM-based models remains an open challenge, primarily because LLMs are mostly trained on text-based tokens and are often not designed to natively process or fuse non-textual inputs.

Beyond technical challenges, it is also crucial for models to address concerns regarding patient trust. Current LLMs are not specifically trained on psychological or psychiatric knowledge comparable to the clinical experience of trained professionals. As a result, the responses generated by LLMs may lack authenticity and, more critically, may exhibit hallucinations—a particularly serious issue in the context of depression recognition. Techniques such as knowledge injection, as discussed by Martino et al. (2023), have been proposed to mitigate this limitation by integrating domain-specific information into LLMs. However, such methods have not yet been widely implemented in LLM-based models for depression recognition, leaving a considerable gap in ensuring both reliability and clinical validity.

Therefore, our paper proposes a novel approach to address the aforementioned research gaps:

- To the best of our knowledge, this is the first approach to directly apply LLMs for spoken language to the field of depression recognition.
- We introduce a pipeline that injects professional psychological knowledge into LLMs and demonstrates its effectiveness through empirical evaluation.
- Our proposed pipeline considerably outperforms baseline models on the DAIC-WOZ dataset.

The remainder of this paper is organized as follows. Section 2.1 reviews related work in depression recognition, particularly focusing on DNNs and LLMs in a fusion of multiple streams. Section 2.2 describes the dataset used in our study, including its composition and its pre-processing. Section 2.3 details our proposed multimodal LLM-based pipeline and the methodology for injecting psychological knowledge into the model. Section 3 presents the experimental results, comparing our approach against established baselines. Finally, Section 4 concludes the paper and outlines potential directions for future research.

2 Materials and methods

2.1 Related work

2.1.1 Text-based depression detection

In recent years, Natural Language Processing (NLP) methods have increasingly utilized deep language models to identify depression symptoms in text. This approach is grounded in clinical practices, where mental health professionals often assess linguistic cues—such as expressions of hopelessness, self-deprecation, or withdrawal—to diagnose depression. In AVEC 2013, Valstar et al. (2013) incorporated the text modality for the first time, providing ASR transcripts alongside audio and video for multimodal depression detection and emotion recognition. Further, Ogunleye et al. (2024) applied multiple hybrid models to two social media datasets, each involving binary classification tasks distinguishing between "depressed" and "not depressed" posts. Their combination of Sentence-BERT and an ensemble model achieved F1 scores of 69% and 76% on the respective datasets, demonstrating that incorporating lexicon-based sentiment indicators can enhance the performance of text-based models. This demonstrates that incorporating lexicon-based sentiment indicators can enhance the performance of text-based models. Similarly, Sivamanikandan et al. (2022), using a social media dataset from the Language Technology for Equality, Diversity, and Inclusion (LT-EDI) 2022 task published by the Association for Computational Linguistics (ACL), trained several transformer models such as DistilBERT, ALBERT, and RoBERTa. Posts in the dataset were categorized as "not depressed," "moderately depressed," or "severely depressed". Among the tested models, RoBERTa performed best, achieving an overall F1 score of 0.457 in the three-class problem, illustrating the effectiveness

of transformer architectures for text-based depression classification tasks. These studies are good examples of the current trend in research, in which researchers are increasingly focused on fine-tuning pretrained language models with labeled text. Previous studies also show that combining deep learning approaches with linguistic markers or psycholinguistic lexicons can substantially improve performance (e. g., Lyu et al., 2023). Kathan et al. (2022b) proposed other format of text-based features such as behavioral activation for depression scale-short form (BADSSF), the center for epidemiologic studies depression scale (CESD), or the personality dynamics diary (PDD). Overall, the trend in text-based depression detection has shifted toward transformer-based models, often emphasizing specific lexical indicators or sentence patterns related to emotional states.

2.1.2 Audio-based depression detection

Depression symptoms can also manifest in vocal expression, prompting researchers to fine-tune pretrained speech models to capture acoustic features. For instance, individuals experiencing depression often exhibit paralinguistic characteristics such as reduced pitch variability, slower speech rate, and longer pauses. These features reflect the low arousal and negative emotional states often associated with depression. In AVEC 2013 Valstar et al. (2013), the first audio-based depression detection challenge, incorporated the audio modality as a key component for depression detection. Moreover, Huang et al. (2024) applied wav2vec 2.0 to the AVEC2017 dataset to extract audio features, achieving 96.5% accuracy in binary depression classification. This highlights the capability of such models to learn high-quality representations without requiring complex processing. Mallol-Ragolta et al. (2024) applied multi-triplet loss-based models for categorical depression recognition with four acoustic features. In addition, classical acoustic analysis remains effective. For example, Berardi et al. (2023) extracted voice pathology features from recordings of picture descriptions and used them to train SVM classifiers. A third-degree polynomial SVM achieved over 92% accuracy across all tasks. Their study identified articulatory precision, pause frequency, and speech variability as the most influential features. These represent two primary approaches in audio-based depression detection: (1) fine-tuning deep models like wav2vec and (2) extracting and classifying key acoustic features using traditional methods like SVM. Both approaches have proven effective for this task.

2.1.3 Multimodal depression detection

Multimodal models have been introduced into depression detection to combine textual, audio, and visual information, enabling richer feature representation than unimodal approaches. A prominent benchmark used in this field is the AVEC 2017 "Real-life Depression and Affect Recognition" challenge, which provides video, audio, and transcript data from interviews. Ringeval et al. (2017) introduced the according dataset with including PHQ-8 score regression. The baseline model for depression severity estimation contains text, audio, video, and combined audio and video models, serving as a reference for future work. Several studies have built upon this benchmark to improve it. Sadeghi et al. (2024) extracted textual and facial-expression features using a LLM and a vision model, combining them to predict PHQ-8 scores. Their multimodal model slightly outperformed the text-only version in terms of mean squared error (MSE). Meanwhile, Min et al. (2023) collected annotated YouTube vlogs and conducted statistical analysis to highlight differences in depressive vs non-depressive videos. Their model learnt from both audio and video cues and achieved an F1 score of 77% on the vlog dataset. Additionally, He et al. (2022) proposes a novel multimodal dataset collected from phone sensors, including phone calls, phone usage, and user activity. Similarly, Kathan et al. (2022a) utilizes mobile sensors to collect multimodal data. These studies demonstrate that integrating multiple modalities—whether text with facial expressions or audio with video—can enhance model performance in depression recognition tasks.

2.1.4 LLM-based depression recognition

The emergence of LLMs has brought major advancements for many application scenarios of AI including depression recognition, outperforming conventional non-large DNNs, as were used before the rise of large models. Schuller et al. (2024) show that LLMs have emotional emergence. Additionally, Shin et al. (2024) demonstrated the effectiveness of LLMs by prompting GPT-3.5 and GPT-4 with 428 diaries from 91 users to assess depression risk. With simple prompt engineering and minimal fine-tuning, the model achieved an accuracy of 90.2% on binary depression classification, where a PHQ-9 score greater than 10 was considered indicative of depression. Notably, the fine-tuned GPT-3.5 outperformed its zero-shot untuned counterpart, underscoring the potential of LLMs when adapted properly. Expanding on this, Liu Z. et al. (2024) introduced "EmoLLMs," a series of LLMs fine-tuned for affective tasks. Trained on a multi-task emotional dataset, EmoLLMs surpassed GPT-4 on standard benchmarks, further showing how LLMs can be tailored to emotional understanding. To assess how closely LLMs resemble human performance, Zhang et al. (2024) compared models like ChatGPT, Claude, and Bing Chat on sentiment intensity tasks. Results showed that GPT-4 achieved scores comparable to humans, suggesting that LLMs can replicate or even exceed human-level emotional judgement. Together, these works indicate that LLMs represent a major leap forward in the field, with vast potential when appropriately fine-tuned.

2.1.5 Knowledge-based injection into LLMs

Building on the strong baseline performance of LLMs, recent research has explored the integration of psychological knowledge to further improve mental health inference. Li et al. (2025) surveyed the effectiveness of continuous learning, where knowledge-based injection served as one potential approach. Specifically, Abbasi et al. (2024) introduced "PsychoLexLLaMA", an LLM designed for psychological assessment. Trained on the PsychoLex Q&A dataset, this model learnt specialized psychological knowledge and outperformed other LLMs in reasoning tasks related to psychology, highlighting the benefits of domain-specific knowledge injection. In another example, Lan et al. (2024) proposed DORIS, a depression recognition system that incorporates clinical diagnostic knowledge. The model first used an LLM to identify social

media posts containing DSM-related expressions, then generates emotional summaries and estimates emotional intensities. These outputs are fed into a conventional classifier, resulting in improved performance for binary depression classification (depressed vs control) compared to standard models. Similarly, Tank et al. (2024) demonstrated that encoding questionnaire knowledge into prompts enhances LLM effectiveness. Their system for PHQ-8 scoring uses structured prompts based on depression symptoms and a two-shot classification setup. They found that embedding relevant questionnaire knowledge increased prediction accuracy. Together, these studies reveal that enriching LLMs with psychological or diagnostic knowledge can further elevate their capability in depression recognition tasks.

2.2 Dataset and preprocessing

2.2.1 Experimental dataset

The dataset used in this study is DAIC-WOZ (Gratch et al., 2014). The interviews feature a patient interacting with a virtual human interviewer named Ellie (as indicated in the transcript files). The dataset includes audio recordings of complete conversations and their corresponding transcripts, which specify the speaker, the spoken content, and the start and end times of each utterance. Although the dataset also contains facial feature data, this was not utilized in the current research.

DAIC-WOZ consists of 189 samples, also referred to as 189 participants, pre-divided into training, validation, and test sets, containing 107, 35, and 47 samples respectively. Additionally, a metadata table is provided, which includes patient IDs, binary depression labels, PHQ-8 scores, and dimensional depression scores (the latter were not used in this study).

2.2.2 Data preprocessing

To enable the LLM to better interpret and extract features from the audio data, the recordings were segmented to accommodate the limited context window of LLMs. Using the transcript file timestamps and speaker annotations, each sentence spoken by a participant was isolated. Every five consecutive utterances were then merged into a single audio file, ensuring that no audio segments from different participants were combined. Given there is a "speaker" value in the dataset, corresponding transcript data was filtered to include only the speech of the participant. The sentences were merged in the same manner as the audio: every five sentences were combined into one element in the new set, subsequently merging the elements into a CSV file. Some additional metadata information values were eliminated in this file, keeping only the merged text content and the associated participant ID. Note that the dataset only provides a single sentence-independent PHQ-8 score per participant, which we therefore take as the target for all sentences from the same speaker.

After preprocessing, a total of 6,556 audio files were generated, each with a corresponding transcript entry, as both were segmented using the same criteria. Each file obtains its label in form of the participant's PHQ-8 score. During the final prediction stage, the model generates a score for each individual audio segment. These segment-level predictions are then averaged to produce the overall PHQ-8 score for the corresponding participant.

2.3 Model pipeline

We have organized our model pipeline as illustrated in Figure 1. We selected the acoustic information as the primary input. The acoustic signal is directly available during clinical interviews, whereas text-based methods require an additional transcription step such as by automatic speech recognition. Nevertheless, the spoken text remains valuable for providing clear and explicit content regarding the subject's verbal expressions. Therefore, we incorporate both types of information to complement each other and enhance overall performance.

We split our LlaMA fine-tuning with two main phases, namely knowledge injection and PHQ-8 score prediction.

2.3.1 Psychology knowledge injection

To address the lack of domain-specific knowledge in LLMs, we design a learning process that mirrors human cognitive development. Specifically, we aim for the LLM to read and comprehend psychological knowledge in a manner similar to how humans study and internalize the information. Inspired by the work of Abbasi et al. (2024), we extract question-answer pairs from psychological sources and prompt the LLM to generate responses accordingly. In contrast to Abbasi et al. (2024), our approach introduces more structured and comprehensive question types to facilitate deeper understanding. Drawing from principles of human learning outlined in Novak and Gowin (1984), effective learning requires understanding what the knowledge is, why it matters, how it is applied, and how it connects with prior knowledge to foster critical thinking and practical use. In line with this, we design six distinct types of questions centered on depressive disorders: (1) the definition of the disorder, (2) the rationale for diagnosing it, (3) common symptoms or manifestations, (4) extended or related knowledge, and (5) critical thinking questions. This structured framework ensures that the LLM not only memorizes facts but also develops a more nuanced and applicable model of depression.

We selected the World Health Organization's official medical classification website[1] as our primary knowledge source due to its authoritative content, which helps reduce the risk of incorporating low-quality or misleading information that could contribute to hallucinations in LLMs. To focus specifically on depression-related knowledge, we extracted a subset of entries by filtering disorder titles using relevant keywords (e. g., anxiety, depress*, mood, stress, chronic, isolation), resulting in a total of 123 samples. We then employed the DeepSeek-V3 model from Liu A. et al. (2024) to generate structured question-answer pairs based on the provided content. Specifically, we used the following prompt to guide the generation process:

> Construct Q&A sets based on one [paragraph] I give you.
> (1) The 10 Q&A sets should be about the key definitions mentioned in the paragraph. The Q&A set should contain no

1 https://icd.who.int/browse/2024-01/mms/en

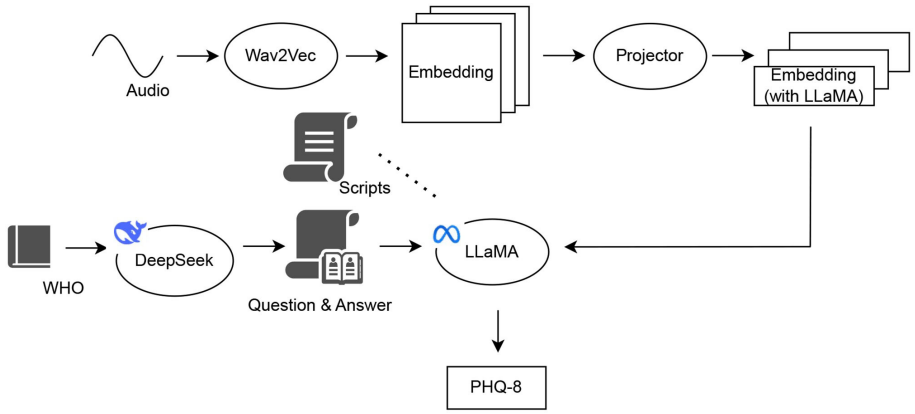

FIGURE 1
Two-Stage Pipeline of the Large Models considered. Our proposed framework consists of two key stages. In the first stage, we leverage the DeepSeek model to extract question-answer pairs from authoritative psychology texts, such as disease definitions and clinical descriptions from the World Health Organization (WHO). These extracted pairs, along with transcript data, are used to pretrain the LLM through a process of knowledge injection, enhancing its domain-specific understanding. In the second stage, we process audio inputs using a feature-extraction DNN, Wav2Vec to obtain their original embeddings. These embeddings are then projected via a feedforward network to align them with the LLM's hidden space. Once the audio and text modalities are integrated, the LLM, LLaMA is used to predict the depression level.

> extra knowledge. (2) The 10 questions in these sets should be "why" questions. The Q&A set should contain no extra knowledge. (3) The 10 Q&A are about the phenomena that may occur on people with such disorder. The Q&A set should contain no extra knowledge. (4) The 5 Q&A sets should be completely based on extended knowledge which is not mentioned in the [paragraph], but should also be considered important about such disorder. (5) The five Q&A sets should show critical thinking. The Q&A set should contain no extra knowledge.
>
> The entire conversation should contain English only. The message you reply must follow the exact format in the [example], do not add any extra " or other marks at the beginning or the end of your question or answer.
>
> [example]:
> question: This is the first question you construct.
> answer: This is the first answer you construct.

This process resulted in a total of 4,920 question-answer pairs. We then provided the questions as prompts to our LLM, LLaMA (Touvron et al., 2023),[2] and employed supervised learning to train the model to generate corresponding answers. Through this process, we aim to effectively inject psychological knowledge into the model, enabling it to better understand and respond to depression-related content.

2.3.2 Multi-stream training

To train the model with multi-stream features, we trained the LLM on text features first using prompts:

> Transcripts:[Transcript], PHQ Score:

This enables the LLM to learn text features first.

Next, to train with audio features and align them with the text-based embedding space of the LLM, we aim to project the audio representations into a shared latent space of the LLM. Inspired by SALMONN, a speech-based LLM proposed by Tang et al. (2023), we adopt a similar pipeline but with a modification: instead of using Whisper (Radford et al., 2023) to extract audio features, which is primarily designed for speech recognition tasks, we utilize Wav2Vec 2.0 (Baevski et al., 2020). We selected Wav2Vec 2.0 due to its stronger capacity for capturing a broader range of audio features, which are more relevant for understanding emotional cues and prosodic elements associated with depression. These have been formulated below.

$$r = \text{Wav2Vec2}(raw_{audio}) \tag{1}$$

$$Emb_{audio} = \text{feedforward}(r) \tag{2}$$

$$\text{PHQ-8} = \text{Linear}(\text{LLaMA}(Emb_{audio})_{-1}), \tag{3}$$

where r is in shape of $R^{s\times d_a}$, with s being extracted audio length by Wav2Vec2 and d_a being the hidden dimension, and Emb_{audio} is in shape of $R^{s\times d_t}$, with d_t being the hidden dimension of LLaMA. In this manner, the audio features are effectively mapped and projected into the embedding space of the LLaMA model, enabling seamless integration with its text-based representations. This allows LLaMA to learn meaningful representations for the audio modality as well. Finally, we utilize the last hidden layer embedding and use a linear layer to predict the PHQ-8 score.

2.3.3 Experiments

To ensure reproducibility, we fixed the random seed to 42 using the *random* libraries within the respective *Numpy* and *torch* libraries in Python. Due to limited GPU resources, we employed Low-Rank Adaptation (LoRA) (Hu et al., 2022) with Distributed Data Parallel (DDP) for efficient fine-tuning of the

2 https://huggingface.co/meta-llama/Llama-2-7b-hf

TABLE 1 LLaMA hyper-parameters.

Parameter	Value
r	8
lora_alpha	16
lora_dropout	0.1
target_modules	"q_proj", "v_proj"
GPU	Tesla V100-PCIE-32GB

TABLE 2 Evaluation on DAIC-WOZ test set for different models.

Model	MAE	RMSE
AVEC2016 (audio) (Valstar et al., 2016)	5.72	7.78
LSTM (Afzal Aghaei and Khodaei, 2023)	5.7	6.59
Random forest (Afzal Aghaei and Khodaei, 2023)	5.71	6.79
Ours(audio)	5.373	6.733
Ours(text)	6.342	8.891
Ours(audio + text)	5.356	6.713
Ours(text + knowledge injection)	**5.354**	**6.429**
Ours(audio + knowledge injection)	5.356	6.713
Ours(audio +text+ knowledge injection)	5.356	6.713

The font value is the best performing score.

LLM. The detailed hyperparameter settings used during training are summarized in Table 1.

3 Results

We conducted a series of experiments, and the results are presented in Table 2 including ablation studies to measure the impact of certain components of our method. To evaluate model performance, we used Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as the primary metrics. Additionally, the final prediction scores are calculated by each participant across multiple clips derived from the same individual.

Our pure audio-based method shows a clear improvement over the baseline, which is only available for audio in the AVEC 2016 challenge (rather than also for text or audio + text) (Valstar et al., 2016). This demonstrates that the audio features in the model contribute to a better performance, suggesting that audio signals provide valuable information for the task, outperforming the baseline in both the MAE and RMSE metrics.

Additionally, knowledge injection has a noticeable impact on LLM depression prediction. It lowers both MAE and RMSE marginally compared to the plain audio and text model. The MAE drops from 5.373 to 5.356 for audio, from 6.342 to 5.354 for text, and RMSE drops from 6.733 to 6.713 for audio, from 8.891 to 6.429, for text. The performance improves more for text than for audio. This suggests that the text-based knowledge injection has a stronger effect, likely because the knowledge is provided in a textual form, which aligns better with the text modality.

When combining both audio and text with knowledge injection, the performance is close to text with knowledge injection, and pure text with audio. While this suggests the potential for combining the information streams, it also highlights that the combination may not always result in substantial performance gains, possibly due to the fact that the learning ability of LLMs is biased toward audio. Additionally, it may be that the LLM's learning ability is limited by small number of trainable parameters from LoRA. However, the combination still performs well and indicates that both modalities can be useful when leveraged appropriately.

Taking another perspective, we investigate whether the knowledge injection works as expected. The procedure involves instructing PsycholexLLAMA to generate fifty questions. The prompt used is relatively straightforward.

> Regard yourself as a teacher teaching about depression detection, and you are trying to test how well your students understand this knowledge. Generate num_questions specific questions about psychological assessment of depression to evaluate whether the students know how to assess depression. Number them as 1., 2., 3., etc.

The generated questions were then submitted to two different models for response, the one trained with the Question & Answer pairs and the one without such post-training. Both the questions and the corresponding answers were subsequently passed back to PsycholexLLAMA, which was tasked with scoring each response on a scale from 0 to 10. Specifically, we adopted this prompt:

> Rate the following response to a question about depression on a scale from 0 to 10:
> QUESTION: question
> RESPONSE: response
> Evaluate based on: - Accuracy of information - Sensitivity to mental health issues - Helpfulness of the response - Appropriate tone
> Rate from 0 (completely inappropriate) to 10 (excellent in all aspects). Your rating (just the number):

For each question, the higher of the two scores was recorded, which follows the principle of the pass@2 strategy to measure performance of LLMs proposed by Chen et al. (2021). Finally, the average of these highest scores was calculated to yield the overall performance score. The original model achieved a score of 7.32, whereas our trained model obtained a score of 8.20, indicating that our proposed strategy is effective and that the model has successfully learnt from the psychology knowledge base.

While our strategy demonstrates promising results, it also highlights two key directions for future work. First, although LoRA offers a lightweight and efficient fine-tuning method, its capacity to induce substantial changes in model behavior is limited compared to full-scale training of LLMs. However, fully training an LLM demands significant computational resources, which may not always be feasible. Second, as previously discussed, there is a notable scarcity of psychologically-informed audio data. As a result, the model—despite being equipped with theoretical knowledge derived from text—struggles to effectively transfer this understanding to audio-based applications. Generating such

audio content artificially using AI can often lack authenticity. A potential solution would be the development of clinically annotated audio datasets that pair spoken examples with corresponding textual explanations—such as diagnostic insights based on criteria from the WHO. Such resources could serve as a valuable bridge between theory and practical application in audio-based psychological assessments.

4 Conclusion

In conclusion, our work presented a novel approach to advancing depression detection by leveraging the capabilities of LLMs. By incorporating professional psychological knowledge into the considered LLM through a carefully designed pipeline, we enhanced the model's ability to interpret and evaluate depressive symptoms more effectively. Our empirical results, validated on the DAIC-WOZ dataset, demonstrate that the proposed method substantially outperforms established baselines (around 0.35 on MAE and 1.36 on RMSE), underscoring the potential of LLMs as a powerful tool in mental health assessment. These findings pave the way for future research exploring the integration of domain expertise into multimodal AI systems for clinical applications. In the future, the work may continue to fully train LLMs and to obtain annotations or suggestions from specialists. Additionally, future work should explore training full-scale LLMs and incorporating richer sound descriptions as a form of knowledge injection, in order to better bridge the gap between acoustic features and the text-based knowledge learned by LLMs.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YL: Conceptualization, Software, Writing – original draft, Project administration, Methodology, Formal analysis, Writing – review & editing, Visualization, Investigation, Validation, Data curation, Resources. SS: Software, Data curation, Investigation, Resources, Formal analysis, Writing – original draft. MM: Funding acquisition, Resources, Writing – review & editing, Supervision. BS: Supervision, Funding acquisition, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was partially supported and funded by the Munich Center for Machine Learning and the Munich Data Science Institute.

Acknowledgments

We acknowledge Hanqian Li from Shandong University for providing initial draft experiment codes, Adria Mallol Ragolta from Technical University Munich to discuss ideas with us.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Abbasi, M. A., Mirnezami, F. S., and Naderi, H. (2024). Psycholex: Unveiling the psychological mind of large language models. *arXiv* [preprint] arXiv:2408.08848. doi: 10.21203/rs.3.rs-4920871/v1

Afzal Aghaei, A., and Khodaei, N. (2023). Automated depression recognition using multimodal machine learning: a study on the daic-woz dataset. *Comp. Mathem. Comp. Model. Appl.* 2, 45–53. doi: 10.48308/CMCMA.2.1.45

Amin, M. M., Cambria, E., and Schuller, B. W. (2023). Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of chatGPT. *IEEE Intellig. Syst.* 38, 15–23. doi: 10.1109/MIS.2023.3254179

Ashraf, A., Gunawan, T. S., Riza, B. S., Haryanto, E. V., and Janin, Z. (2020). On the review of image and video-based depression detection using machine learning. *Indon J. Elect. Eng. Comp. Sci.* 19, 1677–1684. doi: 10.11591/ijeecs.v19.i3.pp1677-1684

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv Neural Inf Process Syst.* 33, 12449–12460. doi: 10.48550/arXiv.2006.11477

Berardi, M., Brosch, K., Pfarr, J.-K., Schneider, K., Sültmann, A., Thomas-Odenthal, F., et al. (2023). Relative importance of speech and voice features in the classification of schizophrenia and depression. *Transl Psychiatry*. 13:298. doi: 10.1038/s41398-023-02594-0

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., et al. (2021). Evaluating large language models trained on code. *arXiv* [preprint] arXiv:2107.03374. doi: 10.48550/arXiv.2107.03374

Farruque, N., Goebel, R., Sivapalan, S., and Zaïane, O. R. (2024). Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach. *Lang. Resources Eval.* 58, 1013–1041. doi: 10.1007/s10579-024-09720-4

Gerber, P. D., Barrett, J., Barrett, J., Manheimer, E., Whiting, R., and Smith, R. (1989). Recognition of depression by internists in primary care: a comparison of internist and "gold standard" psychiatric assessments. *J. Gen. Intern. Med.* 4, 7–13. doi: 10.1007/BF02596483

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., et al. (2014). "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, eds. N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, et al. (Reykjavik, Iceland: European Language Resources Association (ELRA)), 3123–3128.

He, X., Triantafyllopoulos, A., Kathan, A., Milling, M., Yan, T., Rajamani, S. T., et al. (2022). "Depression diagnosis and forecast based on mobile phone sensor data," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Glasgow: IEEE), 4679–4682.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*. 1:3.

Huang, X., Wang, F., Gao, Y., Liao, Y., Zhang, W., Zhang, L., et al. (2024). Depression recognition using voice-based pre-training model. *Sci Rep*. 14:12734. doi: 10.1038/s41598-024-63556-0

Kathan, A., Harrer, M., Küster, L., Triantafyllopoulos, A., He, X., Milling, M., et al. (2022a). Personalised depression forecasting using mobile sensor data and ecological momentary assessment. *Front. Digital Health* 4:964582. doi: 10.3389/fdgth.2022.964582

Kathan, A., Triantafyllopoulos, A., He, X., Milling, M., Yan, T., Rajamani, S. T., et al. (2022b). Journaling data for daily PHQ-2 depression prediction and forecasting," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Glasgow: IEEE), 2627–2630.

Lan, X., Cheng, Y., Sheng, L., Gao, C., and Li, Y. (2024). Depression detection on social media with large language models. *arXiv* [preprint] arXiv:2403.10750. doi: 10.48550/arXiv.2403.10750

Li, Y., Milling, M., and Schuller, B. W. (2025). Neuroplasticity in artificial intelligence-an overview and inspirations on drop in & out learning. *arXiv* [preprint] arXiv:2503.21419. doi: 10.48550/arXiv.2503.21419

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., et al. (2024). Deepseek-v3 technical report. *arXiv* [preprint] arXiv:2412.19437. doi: 10.48550/arXiv.2412.19437

Liu, Z., Yang, K., Xie, Q., Zhang, T., and Ananiadou, S. (2024). "Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (New York: ACM), 5487–5496. doi: 10.1145/3637528.3671552

Lyu, S., Ren, X., Du, Y., and Zhao, N. (2023). Detecting depression of chinese microblog users via text analysis: Combining linguistic inquiry word count (liwc) with culture and suicide related lexicons. *Front. Psychiat.* 14:1121583. doi: 10.3389/fpsyt.2023.1121583

Mallol-Ragolta, A., Milling, M., and Schuller, B. (2024). "Multi-triplet loss-based models for categorical depression recognition from speech," in *Proc. IberSPEECH* 2024, 31–35. doi: 10.21437/IberSPEECH.2024-7

Martino, A., Iannelli, M., and Truong, C. (2023). "Knowledge injection to counter large language model (LLM) hallucination," in *European Semantic Web Conference* (Cham: Springer), 182–185.

Min, K., Yoon, J., Kang, M., Lee, D., Park, E., and Han, J. (2023). Detecting depression on video logs using audiovisual features. *Humanit. Soc. Sci. Commun.* 10, 1–8. doi: 10.1057/s41599-023-02313-6

Niu, M., Zhao, Z., Tao, J., Li, Y., and Schuller, B. W. (2022). Dual attention and element recalibration networks for automatic depression level prediction. *IEEE Trans. Affect. Comp.* 14, 1954–1965. doi: 10.1109/TAFFC.2022.3177737

Novak, J. D., and Gowin, D. B. (1984). *Learning How to Learn*. Cambridge: Cambridge University Press.

Ogunleye, B., Sharma, H., and Shobayo, O. (2024). Sentiment informed sentence bert-ensemble algorithm for depression detection. *Big Data Cognit. Com.* 8:112. doi: 10.3390/bdcc8090112

Qassim, S., Golden, G., Slowey, D., Sarfas, M., Whitmore, K., Perez, T., et al. (2023). A mixed-methods feasibility study of a novel ai-enabled, web-based, clinical decision support system for the treatment of major depression in adults. *J. Affect. Disord. Reports* 14:100677. doi: 10.1016/j.jadr.2023.100677

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning* (New York: PMLR), 28492–28518.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., et al. (2019). "AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 3–12.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., et al. (2017). "AVEC 2017: real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17* (New York, NY: Association for Computing Machinery), 3–9.

Robertson, C., Woods, A., Bergstrand, K., Findley, J., Balser, C., and Slepian, M. J. (2023). Diverse patients' attitudes towards artificial intelligence (ai) in diagnosis. *PLOS Digital Health* 2:e0000237. doi: 10.1371/journal.pdig.0000237

Sadeghi, M., Richer, R., Egger, B., Schindler-Gmelch, L., Rupp, L. H., Rahimi, F., et al. (2024). Harnessing multimodal approaches for depression detection using large language models and facial expressions. *NPJ Mental Health Res.* 3:66. doi: 10.1038/s44184-024-00112-8

Sardari, S., Nakisa, B., Rastgoo, M. N., and Eklund, P. (2022). Audio based depression detection using convolutional autoencoder. *Expert Syst Appl.* 189:116076. doi: 10.1016/j.eswa.2021.116076

Schuller, B., Mallol-Ragolta, A., Almansa, A. P., Tsangko, I., Amin, M. M., Semertzidou, A., et al. (2024). Affective computing has changed: the foundation model disruption. *arXiv* [preprint] arXiv:2409.08907. doi: 10.48550/arXiv.2409.08907

Shin, D., Kim, H., Lee, S., Cho, Y., and Jung, W. (2024). Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: Instrument validation study. *J. Med. Internet Res.* 26:e54617. doi: 10.2196/54617

Sivamanikandan, S., Santhosh, V., Sanjaykumar, N., Jerin Mahibha, C., and Durairaj, T. (2022). "scubeMSEC@LT-EDI-ACL2022: Detection of depression using transformer models," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, eds. B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, and P. Buitelaar (Dublin: Association for Computational Linguistics), 212–217.

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., et al. (2023). Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Inform.* 10:10. doi: 10.1186/s40708-023-00188-6

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., et al. (2023). Salmonn: towards generic hearing abilities for large language models. *arXiv* [preprint] arXiv:2310.13289. doi: 10.48550/arXiv.2310.13289

Tank, C., Pol, S., Katoch, V., Mehta, S., Anand, A., and Shah, R. R. (2024). Depression detection and analysis using large language models on textual and audio-visual modalities. *arXiv* [preprint] arXiv:2407.06125. doi: 10.48550/arXiv.2407.06125

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). Llama: Open and efficient foundation language models. *arXiv* [preprint] arXiv:2302.13971. doi: 10.48550/arXiv.2302.13971

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., et al. (2016). "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 3–10.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., et al. (2013). "AVEC 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge* (New York: ACM) 3–10.

Zafar, F., Alam, L. F., Vivas, R. R., Wang, J., Whei, S. J., Mehmood, S., et al. (2024). The role of artificial intelligence in identifying depression and anxiety: a comprehensive literature review. *Cureus* 16:56472. doi: 10.7759/cureus.56472

Zhang, Z., Peng, L., Pang, T., Han, J., Zhao, H., and Schuller, B. W. (2024). Refashioning emotion recognition modelling: the advent of generalised large models. *IEEE Trans. Comp. Soc. Syst.* 11, 6690–6704. doi: 10.1109/TCSS.2024.3396345